Expectation Maximization (wrap-up) and Intro to Probabilistic PCA

Piyush Rai
IIT Kanpur

Probabilistic Machine Learning (CS772A)

Feb 1, 2016

Parameter Estimation with Latent Variables

- Model $p(\mathbf{X}, \mathbf{Z}|\theta)$, observed data $\mathbf{X}$, latent variables $\mathbf{Z}$, model parameters θ
- Recall GMM, $\mathbf{Z}$: cluster assignments, θ: GMM parameters $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$
- Goal: Estimate the model parameters θ via MLE

$$\hat{\theta} = \arg\max_{\theta} \log p(\mathbf{X}|\theta) = \arg\max_{\theta} \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta)$$

- Doing MLE in such models can be difficult because of the log-sum
- *If* we "knew" $\mathbf{Z}$, sum over all possible $\mathbf{Z}$ not needed. Just define "complete data" $\{\mathbf{X}, \mathbf{Z}\}$, and do MLE on the complete data log-lik. $\log p(\mathbf{X}, \mathbf{Z}|\theta)$
- Assumption: MLE on $\log p(\mathbf{X}, \mathbf{Z}|\theta)$ is easy
  - It often indeed is, especially when $p(\mathbf{X}, \mathbf{Z}|\theta)$ is exponential family distribution (or product of exponential family distributions)

Parameter Estimation with Latent Variables

- If MLE on $\log p(\mathbf{X}, \mathbf{Z}|\theta)$ is easy then let's do it!
- Problem: Well, we don't actually know $\mathbf{Z}$, so we are still stuck. ☹
- Solution: Use the posterior $p(\mathbf{Z}|\mathbf{X}, \theta)$ over latent variables $\mathbf{Z}$ to compute the expected complete data log-likelihood and do MLE on *that* objective.

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta} \mathbb{E}[\log p(\mathbf{X}, \mathbf{Z}|\theta)] \\ &= \arg\max_{\theta} \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta) \log p(\mathbf{X}, \mathbf{Z}|\theta)\end{aligned}$$

- But now we have a chicken-and-egg problem: the posterior $p(\mathbf{Z}|\mathbf{X}, \theta)$ over $\mathbf{Z}$ itself depends on the parameters θ

Solution: An Iterative Scheme (EM Algorithm)

Initialize the parameters: θ^{old}. Then alternate between these steps:

- **E (Expectation) step:**
  - Compute the posterior $p(\mathbf{Z}|\mathbf{X}, \theta^{old})$ over latent variables $\mathbf{Z}$ using θ^{old}
  - Compute the expected complete data log-likelihood w.r.t. *this* posterior

$$\mathcal{Q}(\theta, \theta^{old}) = \mathbb{E}_{p(\mathbf{Z}|\mathbf{X}, \theta^{old})}[\log p(\mathbf{X}, \mathbf{Z}|\theta)] = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{old}) \log p(\mathbf{X}, \mathbf{Z}|\theta)$$

- **M (Maximization) step:**
  - Maximize the expected complete data log-likelihood w.r.t. θ

$$\begin{aligned}\theta^{new} &= \arg\max_{\theta} \mathcal{Q}(\theta, \theta^{old}) \quad \text{(if doing MLE)} \\ \theta^{new} &= \arg\max_{\theta} \{\mathcal{Q}(\theta, \theta^{old}) + \log p(\theta)\} \quad \text{(if doing MAP)}\end{aligned}$$

- If the log-likelihood or the parameter values not converged then set $\theta^{old} = \theta^{new}$ and go to the E step.

Why is this doing the right thing?

Illustration: EM for GMM

- Recall that the GMM parameters $\theta = \{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\} = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^K$
- The complete data likelihood

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{n=1}^{N}\prod_{k=1}^{K} p(\boldsymbol{z}_n = k)p(\boldsymbol{x}_n|\boldsymbol{z}_n = k) = \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(\boldsymbol{x}_n|\mu_k, \Sigma_k)^{z_{nk}}$$

- Taking the log, we get:

$$\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\{\log \pi_k + \log \mathcal{N}(\boldsymbol{x}_n|\mu_k, \Sigma_k)\}$$

- E-step computes the expected complete data log-likelihood:

$$\mathbb{E}_{p(\mathbf{Z}|\mathbf{X},\theta)}[\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})] = \sum_{n=1}^{N}\sum_{k=1}^{K} \mathbb{E}[z_{nk}]\{\log \pi_k + \log \mathcal{N}(\boldsymbol{x}_n|\mu_k, \Sigma_k)\}$$

where $\mathbb{E}[z_{nk}]$ is the expected value of z_{nk} under the posterior

Illustration: EM for GMM (Contd.)

- The only expectation we need to compute $\mathbb{E}_{p(\mathbf{Z}|\mathbf{X},\theta)}[\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})]$ is

$$\mathbb{E}[z_{nk}] = \sum_{z_{nk}=\{0,1\}} z_{nk}\, p(z_{nk}|\boldsymbol{x}_n, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = p(z_{nk} = 1|\boldsymbol{x}_n, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_n|\mu_j, \Sigma_j)} = \gamma_{nk}$$

- Thus the expected complete data log-likelihood

$$\mathbb{E}_{p(\mathbf{Z}|\mathbf{X},\theta)}[\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})] = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma_{nk}\{\log \pi_k + \log \mathcal{N}(\boldsymbol{x}_n|\mu_k, \Sigma_k)\}$$

- M-step maximizes the the exp. complete data log-likelihood w.r.t. π_k, μ_k, Σ_k
- The update equations for these will be (shown on the board)

$$\mu_k = \frac{1}{N_k}\sum_{n=1}^{N} \gamma_{nk}\boldsymbol{x}_n, \quad \Sigma_k = \frac{1}{N_k}\sum_{n=1}^{N} \gamma_{nk}(\boldsymbol{x}_n - \mu_k)(\boldsymbol{x}_n - \mu_k)^\top, \quad \pi_k = \frac{N_k}{N}$$

where $N_k = \sum_{n=1}^{N} \gamma_{nk}$ is "effective" num. of examples assigned to k^{th} Gaussian

Why does EM work?

Justification 1

- Consider the log likelihood on "incomplete" data $\mathbf{X}$

$$\begin{aligned}\log p(\mathbf{X}|\theta) &= \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta) = \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \quad \text{(where } q(\mathbf{Z}) \text{ is some distribution)}\\ &\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \quad \text{(using Jensen's inequality for concave log)}\\ \log p(\mathbf{X}|\theta) &\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}|\theta) - \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z})}_{\text{doesn't depend on } \theta} = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}|\theta) + \text{const.}\end{aligned}$$

- If we set $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \theta)$ then the above inequality becomes equality

$$\begin{aligned}\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta) \log \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{p(\mathbf{Z}|\mathbf{X}, \theta)} = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta) \log \frac{\cancel{p(\mathbf{Z}|\mathbf{X}, \theta)}\, p(\mathbf{X}|\theta)}{\cancel{p(\mathbf{Z}|\mathbf{X}, \theta)}}\\ &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta) \log p(\mathbf{X}|\theta)\\ &= \log p(\mathbf{X}|\theta) \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta) = \log p(\mathbf{X}|\theta)\end{aligned}$$

- Thus for $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \theta)$, we have

$$\log p(\mathbf{X}|\theta) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta) \log p(\mathbf{X}, \mathbf{Z}|\theta) + \text{const.} = \mathbb{E}[\log p(\mathbf{X}, \mathbf{Z}|\theta)] + \text{const.}$$

- EM maximizes $\mathbb{E}[\log p(\mathbf{X}, \mathbf{Z}|\theta)]$, a tight lower-bound on $\log p(\mathbf{X}|\theta)$

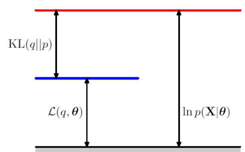
Justification 2

- We can also write the incomplete log likelihood

$$\boxed{\log p(\mathbf{X}|\theta) = \mathcal{L}(q,\theta) + \mathrm{KL}(q||p_z)}$$

where q is some distr. on $\mathbf{Z}$, $p_z = p(\mathbf{Z}|\mathbf{X},\theta)$ is the posterior over $\mathbf{Z}$, and

$$\begin{aligned}\mathcal{L}(q,\theta) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left\{ \frac{p(\mathbf{X},\mathbf{Z}|\theta)}{q(\mathbf{Z})} \right\} \\ \mathrm{KL}(q||p_z) &= -\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left\{ \frac{p(\mathbf{Z}|\mathbf{X},\theta)}{q(\mathbf{Z})} \right\}\end{aligned}$$



(to verify, use $\log p(\mathbf{X},\mathbf{Z}|\theta) = \log p(\mathbf{Z}|\mathbf{X},\theta) + \log p(\mathbf{X}|\theta)$ in the expression of $\mathcal{L}(q,\theta)$)

- Since $\mathrm{KL}(q||p_z) \geq 0$, $\mathcal{L}(q,\theta)$ is a lower-bound on $\log p(\mathbf{X}|\theta)$ for any q

Picture courtesy: PRML (Bishop, 2006)

Justification 2 (contd.)

Recall $\log p(\mathbf{X}|\theta) = \mathcal{L}(q,\theta) + \mathrm{KL}(q||p_z)$. EM can also be seen as:

- With θ fixed to θ^{old}, maximize $\mathcal{L}(q,\theta^{old})$ w.r.t. q

$$\boxed{\hat{q} = \arg\max_q \mathcal{L}(q,\theta^{old})}$$

which is equivalent to making $\mathrm{KL}(q||p_z) = 0$ or setting $\hat{q} = p(\mathbf{Z}|\mathbf{X},\theta^{old})$

(This step makes $\mathcal{L}(\hat{q},\theta^{old}) = \log p(\mathbf{X}|\theta^{old})$; see next slide)

- With $\hat{q}$ fixed at $p(\mathbf{Z}|\mathbf{X},\theta^{old})$, maximize $\mathcal{L}(\hat{q},\theta)$ w.r.t. θ, where

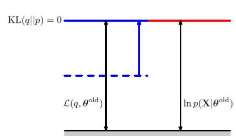
$$\begin{aligned}\mathcal{L}(\hat{q},\theta) &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\theta^{old}) \log p(\mathbf{X},\mathbf{Z}|\theta) - \underbrace{\sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\theta^{old}) \log p(\mathbf{Z}|\mathbf{X},\theta^{old})}_{\text{constant w.r.t. } \theta} \\ &= \mathcal{Q}(\theta,\theta^{old}) + \text{const}\end{aligned}$$

$$\boxed{\theta^{new} = \arg\max_\theta \mathcal{Q}(\theta,\theta^{old})}$$

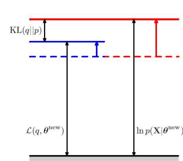
(This step ensures that $\log p(\mathbf{X}|\theta^{new}) \geq \log p(\mathbf{X}|\theta^{old})$; see next slide)

Justification 2 (contd.)

E-step: $\mathcal{L}(q,\theta^{old})$ increases and becomes equal to $\log p(\mathbf{X}|\theta^{old})$, $\mathrm{KL}(q||p_z)$ becomes 0 because we set $q = p(\mathbf{Z}|\mathbf{X},\theta)$



M-step: θ^{new} makes $\mathcal{L}(q,\theta^{new})$ go further up, makes $\mathrm{KL}(q||p_z) > 0$ again because $q \neq p(\mathbf{Z}|\mathbf{X},\theta^{new})$ and thus ensures that $\log p(\mathbf{X}|\theta^{new}) \geq \log p(\mathbf{X}|\theta^{old})$

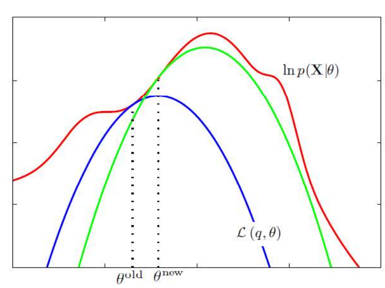


Thus the E and M steps never decrease the log-likelihood $p(\mathbf{X}|\theta)$

Picture courtesy: PRML (Bishop, 2006)

A View in the Parameter Space

- E-step: Update of q makes the $\mathcal{L}(q,\theta)$ curve touch the $\log p(\mathbf{X}|\theta)$ curve
- M-step gives the maxima θ^{new} of $\mathcal{L}(q,\theta)$
- Next E-step readjusts $\mathcal{L}(q,\theta)$ curve (green) to meet $\log p(\mathbf{X}|\theta)$ curve again
- This continues until a local maxima of $\log p(\mathbf{X}|\theta)$ is reached



Some EM Variants

- **Generalized EM:** M step doesn't require maximization w.r.t. θ; even if the new θ just increases the lower bound, we will still converge to a local optima
- **Variational EM and MCMC EM:** If the E step of computing the posterior $p(\mathbf{Z}|\mathbf{X},\theta)$ is intractable, we can use variational Bayes (VB) or MCMC to approximate the posterior
- **Expectation Conditional Maximization:** Parameters are partitioned in groups. M step consists of multiple steps (each optimizing one group of parameters, treating all other groups as fixed)
- **Online/incremental EM:** E step only processes one (or a small number of) observation, computing posteriors/expectations only w.r.t. that minibatch of data. For exponential famility distributions, the sufficient statistics needed in the M step can be easily updated incrementally, leading to simple form of incremental parameter updates. Very useful for scalable inference. See:
  (1) Online EM Algorithm for Latent Data Models (Cappé & Moulines, 2009)
  (2) Online EM for Unsupervised Models (Liang & Klein, 2009)

Next up: Probabilistic PCA and Factor Analysis